Published online 11 November 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D549–D557***
doi: 10.1093/nar/gku1137

CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13C-fluxomics

Zhengdong Zhang[1,†], Tie Shen[2,†], Bin Rui[3,†], Wenwei Zhou[2], Xiangfei Zhou[4], Chuanyu Shang[4], Chenwei Xin[3], Xiaoguang Liu[2], Gang Li[4], Jiansi Jiang[4], Chao Li[2], Ruiyuan Li[4], Mengshu Han[4], Shanping You[4], Guojun Yu[4], Yin Yi[4], Han Wen[3,*], Zhijie Liu[2,*] and Xiaoyao Xie[2,*]

[1]College of Computer Science and Technology, Guizhou University, Guiyang, Guizhou 550025, P.R. China, [2]Key Laboratory of Information and Computing Science Guizhou Province, Guizhou Normal University, Guiyang, Guizhou 563000, P. R. China, [3]School of Life Sciences, Anhui Agricultural University, Hefei, Anhui 230026, P. R. China and [4]School of Life Sciences, Guizhou Normal University, Guiyang, Guizhou 563000, P. R. China

Received August 13, 2014; Revised October 19, 2014; Accepted October 27, 2014

ABSTRACT

The Central Carbon Metabolic Flux Database (CeCaFDB, available at http://www.cecafdb.org) is a manually curated, multipurpose and open-access database for the documentation, visualization and comparative analysis of the quantitative flux results of central carbon metabolism among microbes and animal cells. It encompasses records for more than 500 flux distributions among 36 organisms and includes information regarding the genotype, culture medium, growth conditions and other specific information gathered from hundreds of journal articles. In addition to its comprehensive literature-derived data, the CeCaFDB supports a common text search function among the data and interactive visualization of the curated flux distributions with compartmentation information based on the Cytoscape Web API, which facilitates data interpretation. The CeCaFDB offers four modules to calculate a similarity score or to perform an alignment between the flux distributions. One of the modules was built using an inter programming algorithm for flux distribution alignment that was specifically designed for this study. Based on these modules, the CeCaFDB also supports an extensive flux distribution comparison function among the curated data. The CeCaFDB is strenuously designed to address the broad demands of biochemists, metabolic engineers, systems biologists and members of the -omics community.

INTRODUCTION

Upon entering the post-genome period, the emergence of modern technology has driven a huge wave of -omics research, such as transcriptomics, proteomics, metabolomics, 13C-fluxomics and so on (1,2). As massive amount of data has been accumulated through projects using various -omics, well-known databases have emerged in large numbers for the distribution and comparative analysis of these data (3,4). A number of transcriptome databases have arisen for general and specific purposes, such as the general-purpose Microbial Transcriptome Database and HBT for Human transcriptome (5–7). In addition, there has been a tremendous increase in the number of databases for metabolomics and metabolic pathway data. MetaboLights was constructed for documentation of metabolite structures, their reference spectra and experimental data from metabolic experiments (8). The Human Metabolome Database was built as a database containing detailed information regarding small molecule metabolites found in the human body (9). The famous MetaCyc database provides comprehensive and freely accessible resource of metabolic pathways and enzymes (10). The goal of the BiGG database is to address the need for access to high-quality curated

*To whom correspondence should be addressed. Tel: +86 0851 6702821; Fax:+86 0851 6702822; Email: xyx@gznu.edu.cn
Correspondence may also be addressed to Han Wen. Tel: +86 0551 63710182; Fax: +86 0551 63359188; Email: swhx12@ahau.edu.cn
Correspondence may also be addressed to Zhijie Liu. Tel: +86 0851 6702821; Fax: +86 0851 6702822; Email: liuzj@gznu.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

metabolic models and reconstructions among the systems biology community (11).

Among these -omics, 13C-fluxomics is a new area that involves experimentally quantifying the rates of metabolic reactions within the central carbon metabolism through ^{13}C metabolic fluxomics (12–15). This method enables the resolution of paralleled and reversible reactions (14). Generally, fluxomics delivers a flux estimate describing the metabolic state of a cell at a given time (16). Because metabolic fluxes are the ultimate identifier of a cell's functional state, the result represents the physiological counterpart of its sibling-omics and is supposed to serve as the critical link between genes, proteins, metabolites and the observable phenotype (17). The incorporation of fluxomics as a basal part of the cellular system allows for a deeper understanding of the cellular physiological properties and provides the most comprehensive integration of the different -omics levels (18,19). Recently, 13C-fluxomics has been employed to discriminate between different physiological states, to investigate metabolic responses to genetic manipulations and environmental stress and to monitor the presence of certain metabolic pathways (20–23).

In parallel with other -omics, 13C-fluxomics has documented large amounts of information regarding the flux distribution of numerous organisms (24). Comparative analysis and theoretical simulation of these results prove to be an extraordinarily powerful method for exploring the hidden meanings among numerous fluxomics results (25–28). Comparisons of experimental flux profiling results with those predicated by flux balance analysis have been implemented for normal melanocytes and melanoma cell lines under normoxic and hypoxic conditions, providing a basis for targeting metabolism for the therapeutic benefit of melanoma (26). By comparing the metabolic fluxes of a large number of deletion mutants of *B. subtilis*, Blank *et al*. found that some mutants grew faster than the corresponding wild strain. This result suggests that bacteria may have a regulatory mechanism that ensures that the metabolic network does not function according to maximized biomass production (29). On the basis of a comparative study of flux measurements from nine bacteria and multiobjective optimization theory, Schuetz *et al*. showed that flux states evolve obeying the trade-off between two principles: optimality at one given condition and minimal adjustment between conditions (30). Furthermore, a comparison between the 'dry' flux results and 'wet' flux results has obtained increasing universality with the progressive demand for simulation in metabolic systems biology (31,32).

However, a large fraction of the fluxomics results has been reported in non-standard graphical form (33). Each of the related metabolic networks was unique and incompatible with each other due to the varied definition of the lumped reactions (34). Furthermore, a common measure is expected for calculating similarities or distances between different flux distributions. As a result, unlike other -omes, there is still future potential for a public and user-friendly online tool for the documentation, visualization and comparative analysis of quantitative flux results. To this end, we present the Central Carbon Metabolic Flux Database (CeCaFDB, http://www.cecafdb.org/), which, to the best of our knowledge, is the first professionally designed online tool for fluxomics data regarding the central carbon metabolic systems of microbes and animal cells as a platform for data distribution, visualization and alignment in the systems biology community.

The CeCaFDB encompasses 581 cases of flux distribution and provides a Cytoscape Web-based (35) interactive visualization service for the flux data stored in the database or submitted by individual users. Furthermore, the CeCaFDB utilizes three different algorithms, including a vector-based method, a stoichiometry-based method and a topology-based method to compare/align different flux distributions, supporting a comprehensive and flexible solution for comparative flux analysis. Together with the curated flux data, a comparison service is capable of performing high-throughput correlation analysis and of evaluating the phylogenetic relationships of currently available flux distributions. We hope that the CeCaFDB will benefit from all of the research related to fluxomics and will grow into an influential functional database for fluxomics and systems biology.

DATABASE DESCRIPTION

Data Source

The PubMed reference database was queried using various combinations of keywords, such as '13C', 'metabolic', 'flux' and 'analysis', returning approximately 1000 literature results from 1995 to 2013. Seventy percents of them were automatically excluded due to the absence of quantitative flux distribution information. Then a further 10% were measured using a non-13C method and were also omitted. Ultimately, a total of 118 references were collected as a preliminary source for our database. The diversity of the reactions and network definition, the quantity of experimental data and the required genetic and cultivation knowledge made the assembly of the CeCaFDB both difficult and time consuming. The lumped reactions in these studies were broken down into their original forms, as in the KEGG Reaction Database (36), and the flux value was mapped to its precisely corresponding reaction. Currently, the database encompasses 581 cases of flux distributions from 36 organisms. The fluxomics result is displayed in table format with an interactive graphic representation based on the Cytoscape Web platform. The technical architecture of CeCaFDB was shown in Figure 1. The basal statistics are displayed in Table 1.

Data Content

Each individual case of flux distribution includes several data fields describing the origin of the information, the process measure, a case-specific description and the flux distribution. The 'Experiment name' represents the citation information for the reference. The 'Strain' denotes the specific strain name of the organism if it was mentioned in the reference. The 'Culture medium' identifies the chemical composition of the growth medium. As the name suggests, the 'Carbon source' represents the carbon source in the medium. The 'Growth rate' indicates the average growth rate measure for cases that have a critical parameter determining the flux distribution. The 'Specific rate' includes spe-

Table 1. The basal statistics of all collected flux distributions

Species Total	Reference 118	Condition[a] 581	Flux graph[b] 581	Species	Reference	Condition	Flux graph
Actinobacillus succinogenes	2	5	5	*Mycobacterium tuberculosis*	1	5	5
Agrobacterium tumefaciens	1	8	8	*Penicillium chrysogenum*	1	2	2
Arthrobacter sp.	1	3	3	*Pichia pastoris*	3	7	7
Ashbya gossypii	1	2	2	*Pseudomonas aeruginosa*	1	2	2
Aspergillus nidulans	2	6	6	*Pseudomonasfluorescens*	1	1	1
Aspergillus niger	2	4	4	*Pseudomonas putida*	2	3	3
Bacillus megaterium	2	11	11	*Rhodobacter sphaeroides*	1	1	1
Bacillus subtilis	11	47	47	Rhodopseudomonas palustris	1	2	2
Basfia Succiniciproducens	1	1	1	*Saccharomyces cerevisiae*	14	76	76
Chlorobaculum tepidum	1	2	2	*Scheffersomyces stipitis*	1	4	4
Corynebacterium glutamicum	22	57	57	*Schizosaccharomyces pombe*	1	2	2
Desulfovibrio vulgaris	1	1	1	*Shewanella oneidensis*	1	3	3
Escherichia coli	32	297	297	*Shewanella spp.*	1	11	11
Geobacillus thermoglucosidasius	1	2	2	*Sinorhizobium meliloti*	1	1	1
Geobacter metallireducens	1	2	2	*Synechocystis sp.*	1	1	1
Gluconacetobacter xylinus	1	3	3	*Thermus thermophilus*	1	1	1
Homo sapiens	2	3	3	*Xanthomonas campestris*	1	1	1
Methylobacterium extorquens AM1	1	2	2	*Zymomonas mobilis*	1	3	3

[a]Condition stands for individual flux distribution that might be differentiated from the others by strain type, genotype and cultivation environment.
[b]Flux graph stands for the flux graph visualized by the CytoScape Web based on the corresponding flux values.

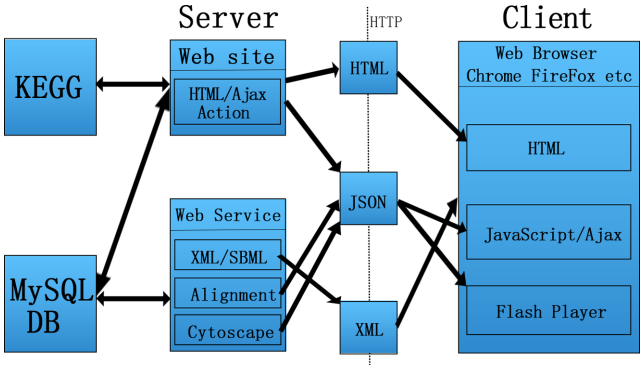

Figure 1. Representation of the system architecture for the CeCaFDB.

cific uptake or secretion rates, available in the references, for carbon source and other primary metabolites with the unit of mmol $g^{-1}h^{-1}$. The 'Case-specific description' contains other information necessary for differentiating cases in a single experiment.

Browse Function

Detailed information regarding where to locate and how to use the browsing tools is provided on the CeCaFDB website (Figure 2). The CeCaFDB provides two entry points for browsing items related to a complete description of a metabolic flux map.

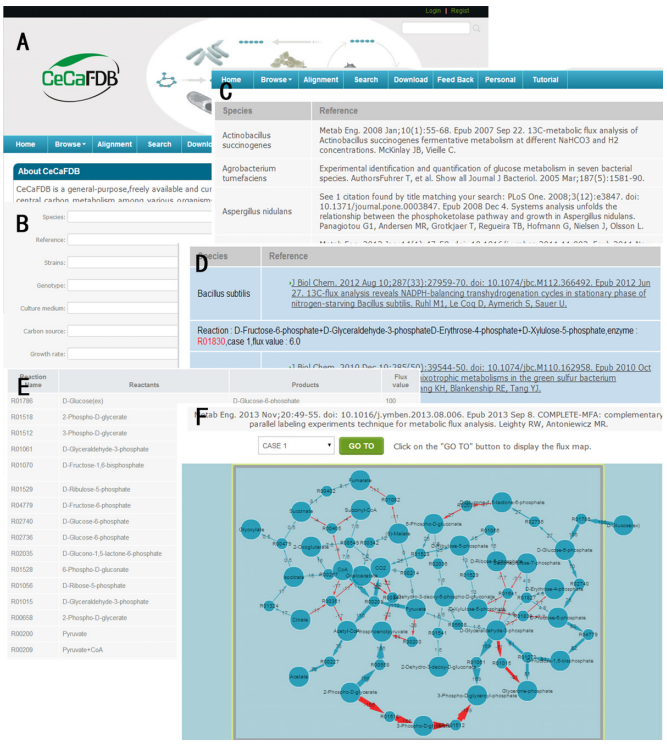

Figure 2. Screenshot montage of the CeCaFDB showing several of its data display and search tools. (**A**) Home page of CeCaFDB. (**B**) Flux distribution browser page. (**C**) Search panel. (**D**) Searching results page. (**E**) Example of flux distribution value. (**F**) Interactive graph representation of flux distribution of *Escherichia coli*.

Browse metabolites and reactions

This entry point is directed to a page including the category information of the metabolites and reactions involved in all the flux maps. On this page, all the metabolites 'belonging to the central carbon metabolism' curated in the database were defined in a metabolite index list since the scope of CeCaFDB was concentrated on central carbon metabolism. Only reactions where the reactants and products belonged to the metabolite index list were included in the database. The reactions not meeting this description were excluded from our database due to the limited amount of literature. This web page contains a total of 66 kinds of metabolites and 76 kinds of chemical reactions constituting the central carbon metabolite systems for all of the collected organisms. For a more detailed introduction to the metabolites and reactions, the URLs were provided in the following column linking to the KEGG COMPOUND (36,37).

Browse flux distribution

This entry is directed to the hierarchy and allows the users to taxonomically browse the database (Figure 2). Upon clicking the button, a summary page is displayed, providing a general description of the organisms and the relevant references. Clicking an organism name will display the whole list of references related to this organism. Users can easily navigate to the flux map pages within the reference (Figure 2). The structure of the flux map page contains several parts. The first part is an interactive graphical representation (see below) of the flux distribution based on the Flash-based Cytoscape Web software. A select box next to a 'Go To' button is located above the graph. To view the graph, you can select a favorite case of flux distribution through the select box and click the 'Go To' button.

The second part is a table representation of the flux distribution. Here, the 'Reaction name' denotes the KEGG code for a specific reaction. There are three types of 'Reaction name' in CeCaFDB. The first is a classical reaction code such as R01002 in KEGG, which denotes an elemental reaction in the central carbon metabolism. The second is labeled as 'transport' that represents all transport across biological membranes, such as malate transport or succinate transport. The third is labeled as 'unknown' that includes all lumped reactions connecting carbon sources to central carbon metabolites. The 'Reaction' displays the chemical equation that occurred in the reaction. The 'Flux value', the core of the database, displays the quantity of the flux value relative to the substrate uptake rate. In the case of multiple substrates, the sum of all the substrates uptake rates is set to 100.

The NADPH and ATP production capacities were recalculated based on the flux values. Where the flux values are missing for certain reactions contributing to NADPH or ATP production, the production rate is shown as N/A. The P/O ratio was set as 2.5 for NADH and 1.5 for FADH2. Since the substrate specificity of isocitrate dehydrogenase and malic enzyme might wobble between NADH and NADPH, a determined ATP or NADPH production rate is rarely available. Alternatively, CeCaFDB provides the maximum and minimum of NADPH and ATP production rates by assuming those enzymes operated exclusively with NADH or NADPH.

The third section of the page contains information for each flux distribution and a link to the original source of the reference. The description information contains subitems, including the 'Strain', 'Culture medium', 'Carbon source', 'Growth rate', 'Specific rate' and 'Case-specific description' (see above for descriptions of their meanings).

Search Function

As with any web-enabled database, the CeCaFDB supports standard text queries. Users can combine multiple search terms in the easy search facility on the homepage, for example, 'coli G6PDH' will generate all of the studies that include the terms 'coli' and 'G6PDH'. Furthermore, it is possible for users to further refine a search using the advanced search, which can be accessed using the SEARCH tab (Figure 2). In this functional module, the search facility offers access to search free text using the underlying data fields, including the experiment title, species name, genetic background, cultivation conditions and protocols, enzymes and flux values. For instance, if the 'Enzyme' field has 'R00402' (reaction number) entered and the 'Flux value' field has '45 and 50' entered, the returned results will only contain flux distributions with enzyme R00402 and a flux value between 45 and 55. Figure 2D shows the resulting page when searching for a flux value between 45 and 55 across the entire CeCaFDB. Each hit contains two segments of information. The first is the species names and the second is the reference name and the matching field, which is emphasized by red color in the search. Users can simply click on the reference title to learn more of the details.

Interactive visualization of flux distribution

The CeCaFDB utilizes the Cytoscape Web API to visualize representations of the metabolic flux maps (35). This interactive visualization tool has its roots in the popular Cytoscape Web platform, but uses Flash technology instead of Java to reduce launch time. It is compatible with any web browser. The latest version of Cytoscape Web works best with up to several thousands of nodes and edges that completely satisfies the requirements of flux map visualization. The flux map is delivered into the Cytoscape Web API using selected parameters, drawing a dynamic graphical display that enables users to move and modify the node and edge properties. Furthermore, the graph can be panned and zoomed in the same layout. In the default configuration, a metabolite is denoted by a blue circle and the reaction by a blue eclipse (Figure 2). A light blue edge denotes a forward flux, whereas a light red edge represents a backward reaction. The edge width is proportionate to the flux value, which is also displayed on the graph with black letter. The boundary between the compartmentation of the organelle and the intra/extracellular environment is represented by closed brown lines, and the compartmentation names are displayed. The chemical transport between different compartmentations and the substrate uptake are embodied by the layout to ensure a meaningful picture.

Flux Comparison Function

Vector-based comparison. An intuitive representation of the flux distribution takes the form of a vector V whose element is the value of each reaction in the concerned metabolic network (12). In such a way, the comparison and analysis between the flux distributions can be accessed by the angular cosine of different distribution vectors

$$\text{Similarity} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2}\sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

A and B represent different vectors and n represents the dimension of the vectors.

Stoichiometry-based comparison. The vector treatment of the flux distribution comparison takes different reactions into equal consideration without losing information regarding the relative weight of the reactions. However, the reaction's impact on the material flow of the metabolic network varies based on the diverse number of consumed reactants and different amounts of product, namely the stoichiometric constraints. Hence, a more appropriate representation of the flux distribution and a description of the flux comparison should incorporate the chemical stoichiometric constraints into the flux value. With the intention of describing the flux distribution with stoichiometric information, the flux matrix F was created by multiplying the stoichiometric matrix by the flux vector as follows:

$$F = S'V' \quad (2)$$

where S stands for the stoichiometric matrix of a metabolic network. The flux matrix F incorporates the chemical stoichiometric constraints into the description of the flux value and is suitable for further comparison of the flux distribution. In the flux matrix framework, similarity between the flux distributions can therefore be defined as follows:

$$\text{Similarity} = \frac{AB'}{\sqrt{\sum_{i,j=1}^{l,m} (A_{ij})^2}\sqrt{\sum_{i,j=1}^{l,m} (B_{ij})^2}} \quad (3)$$

where A and B stand for the different flux matrixes. l and m are the dimensions of the matrixes. This similarity measurement has assimilated information regarding both the flux values and stoichiometric coefficients; therefore, it might be a better approximation of the similarity of the flux distribution.

Topology-based comparison. However, even the stoichiometry-based method cannot represent solid consideration of a metabolic flux comparison. The metabolic flux is flowing through a metabolic network with a specific topological structure for a particular function. Because the topology and function relationship is of great importance for all the metabolic networks (38), it is necessary for the comparison tool to be powerful enough to take into account both the network architecture similarities and the node similarities (39,40). To achieve this objective, a pathway alignment algorithm was modified from the one proposed by Zhengping *et al.* (40) and was combined with the flux value weight to adapt to flux distribution comparisons. In short, the metabolic flux distribution comparison problem was based on a weighted directed graph alignment representation. The similarity calculation between the two metabolic flux maps was then transformed into an optimal weighted alignment between the two graphs.

A metabolic flux distribution with l metabolites and n reactions (see example in Supplementary Figure S2) can be treated as a weighted directed graph $G(N,E,W)$ whose nodes N correspond to the reactions and whose edges E connect with the nodes if the product of one reaction serves as the substrate of the other (40). Before formulating the flux comparison problem with a graphical representation, we should first parameterize the weighted adjacent matrix and node-similarity matrix.

Weighted adjacent matrix. Let m_{ij} denote the summed stoichiometry coefficients of the compounds shared by the products of the i-th reaction and the reactants of the j-th reaction. The $m_{ij} = \delta\left(\left|\sum_{ki} S_{ki}\right| + \left|\sum_{kj} S_{kj}\right|\right)$. δ equals 1 when $\sum_{ki} s_{ki} \geq 0$ and δ equals -1, when $\sum_{ki} s_{ki} < 0$. The value of m_{ij} is set to zero when no common compounds are shared by the two reactions. Let v_i denote the flux value of the i-th reaction. The weight of the graph is introduced by constructing the weighted adjacent matrix $W\{w_{ij}\}$, which is elaborately designed to incorporate both the flux value and stoichiometry. The determination of the elements of W is described by Equation (4) and a simplified example is displayed in Supplementary Figure S2.

$$w_{ij} = \begin{cases} \frac{v_i}{\sqrt{\sum_{i=1}^{n}(v_i)^2}}, i = j \\ \frac{m_{ij}}{\sqrt{\sum_{i,j}^{n}(m_{ij})^2}} = \frac{-m_{ji}}{\sqrt{\sum_{i,j}^{n}(m_{ij})^2}}, i \neq j \end{cases} \quad (4)$$

Node similarity function

There have been several similarity functions proposed to produce a similarity score between a pair of reactions or enzymes, such as functions based on the similarity of amino acid sequences or information content regarding an enzyme class hierarchy (41–43). In our model, a function was based on the probability that two enzymes are the same in the enzyme hierarchy. The enzyme hierarchy is the hierarchy constructed with the EC numbering system (e.g. [3:2:2:1], [3:2:1]). For the two enzymes u and v, a common upper class is defined as the enzyme class h_{uv}, which is the lowest class in the upper classes of enzymes on the enzyme hierarchy. For the same enzymes, their common upper class is their enzyme class. For example, [2:2:3] is the common upper class between [2:2:3:4] and [2:2:3:5]. The $C(h)$ expresses the number of elements for all of the enzymes whose classes are included under the enzyme class h. A similarity function between e_u and e_v is defined as follows:

$$S_{uv} = \frac{1}{C(h_{uv})} \quad (5)$$

which provides a normalized similarity function that incorporates values in the interval [0, 1].

In our algorithm, given two metabolic flux maps $G_1 = G_1(N_1, E_1, W_1)$ and $G_2 = G_2(N_2, E_2, W_2)$, where $N_1 =$

$\{n_1^1, n_2^1, \ldots, n_o^1\}$ and $N_2 = \{n_1^2, n_2^2, \ldots, n_p^2\}$, the weighted adjacent matrixes of G_1 and G_2 are $A = (a_{ij})_{oxo}$ and $B = (b_{ij})_{pxp}$, as constructed in Equation (5). As suggested by Zhengping *et al.* (40), the matching between nodes $(n_i^1, n_k^1) \in N_1$ and $(n_j^1, n_l^1) \in N_1$ and between edges $(n_i^1, n_k^1) \in N_1$ and $(n_j^1, n_l^1) \in N_1$ is represented by the binary variables x_{ij} and y_{ijkl} respectively, as follows:

$$x_{ij} = \begin{cases} \text{one, if } n_i^1 \in N_1 \text{ matches } n_j^2 \in N_2 \\ \text{zero, otherwise} \end{cases}$$

$$y_{ijkl} = \begin{cases} \text{one, if } n_j^1 \in N_1 \text{ matches } n_j^2 \in N_2 \text{ and } n_k^1 \in N_1 \text{ matches } n_j^2 \in N_2 \\ \text{zero, otherwise} \end{cases}$$

Obviously, each $X = \{x_{ij}\}$ and $Y = \{y_{ijkl}\}$ determine the local alignment between two flux maps G_1 and G_2. The similarity between the two flux maps G_1 and G_2 according to a given alignment matrix X of nodes, is thus calculated as a sum score including both the node and edge matching scores in an objective function, as in Equation (7), which is similar to the form of the inner product of the flux vector.

$$max_X f(G_1, G_2) = \lambda \sum_{i=1}^{o}\sum_{j=1}^{p} S_{ij} a_{ii} b_{jj} x_{ij} + (1-\lambda) \sum_{i=1}^{o}\sum_{j=1}^{p}\sum_{k=1}^{o}\sum_{l=1}^{p} a_{ik} b_{jl} y_{ijkl}$$

$$s.t. \begin{cases} \sum_{j=1}^{p} x_{ij} \le 1, i = 1, 2, \ldots, o \\ \sum_{i=1}^{o} x_{ij} \le 1, j = 1, 2, \ldots, p \\ x_{ij} \ge y_{ijkl}, i, k = 1, 2, \ldots, o; j, l = 1, 2, \ldots, p \\ x_{kl} \ge y_{ijkl}, i, k = 1, 2, \ldots, o; j, l = 1, 2, \ldots, p \\ x_{ij} = 0/1, i, k = 1, 2, ..., o; j, l = 1, 2, ..., p \\ y_{ijkl} = 0/1, i, k = 1, 2, \ldots, o; j, l = 1, 2, \ldots, p \end{cases} \quad (7)$$

The first two constraints ensure that the relationship between two nodes is a one-to-one correspondence or involves no matches. The third constraint implies the integer constraint for the variable x. In this framework λ is a scaled parameter between 0 and 1 that is aimed at reaching a compromise between the node (flux value) and edge (stoichiometry) score. The structure of this objective function provides a normalized similarity function that incorporates values from the interval $[-1, 1]$. A pair of identical flux maps yield a similarity function of 1, whereas a pair of flux maps with identical structures and reversed reactions yield a similarity function of -1.

Alignment tab

Based upon the three algorithms mentioned above, the CeCaFDB's Alignment Tab provides four different comparison tasks: 'Vector-based similarity', 'Stoichiometry-based comparison', 'Enzyme Topology-based similarity' and 'Topology-based similarity'. For the vector- and stoichiometry-based tasks, a calculation is carried out on the shared reaction set between the two flux distributions. The 'Enzyme topology-based similarity' is implemented using integer programming by taking λ as 1. The 'Topology-based similarity' should be launched with a user-designated parameter-vertex to the edge score balance (λ value), which determines the relative impact of the vertex and edge scores on the final similarity score. This can be fine-tuned according to the user's need. We have implemented a proposed integer-programming algorithm using the YALMIP and Gurobi softwares. The 2D X and 4D Y were transformed into 1D data and compressed into one single array to save memory requirements and to adapt to the requirement of the optimization package.

By clicking the ALIGNMENT tab, user will enter the flux comparison page. On this page, there are three select boxes for 'species', 'reference' and 'flux distribution'. Through the use of these three boxes, the specific flux distributions to be compared can be selected by the user. After that, clicking the SELCET button will transfer the selected case to the SELECTED box. Repeating this process will add all the desired cases to the SELCETED box. The cases can be deselected by clicking the REMOVE button. Clicking the COMPARISON button will initiate a pairwise similarity analysis on the flux distributions in the SELECTED box except that the 'Topology-based similarity' currently receives the alignment job only between two flux distributions.

The comparison results from the latter two methods contain a similarity score, a *P*-value and an URL link to the graphical representation of the alignment solution. The statistical *P*-value was calculated with a Monte Carlo permutation test, in which the same comparison was executed against a 100-flux map with random sampled values and with the *P*-value obtained by counting the fraction of the flux distributions containing alignments that received higher scores (40). The legends in the graphical representation of the flux comparisons are similar to those in the flux visualization, except that the olive line connecting the enzymes represents the relationship match between the two flux maps, whereas the line width was proportionate to the contribution factor from each enzyme pair to the ultimate similarity score. In addition, on this graphical representation page, CeCaFDB offers a downloadable txt file with tabular form information about the matched reactions, conserved metabolic pathways, gaps in the network and solely inserted reactions. Relying on this alignment tab, CeCaFDB supports flux map query tasks against all the curated flux map data through similarity comparisons after the data have been submitted to CeCaFDB.

Data Download and Submission

Users may download the CeCaFDB data containing metabolic networks and flux distribution vectors. Upon clicking the Download tab, a user will open a download page. Clicking the corresponding reference name will display the links for the corresponding flux distribution that is stored in an Excel file.

The CeCaFDB receives submissions of the flux distribution data. The submission of data requires an account within the CeCaFDB, which can be obtained through online registration. For submissions, the flux data should be formatted in a template file on the submission page, which is similar to the input template for the VANTED software (44). The input file should contain the chemical equations and E.C numbers (or KEGG reaction number) of the metabolic network and the flux distribution values. Particularly important are the lumped reactions of the uploaded

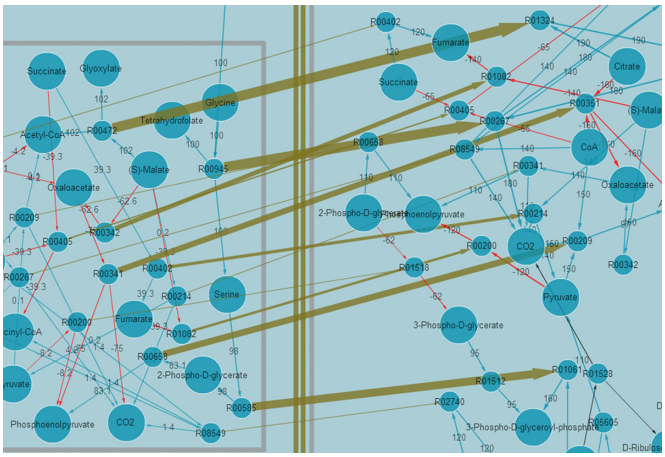

Figure 3. Example of an alignment between the flux distributions of *M. extorquens* AM1 and *E. coli*. The submap containing the enzymes L-Malate glyoxylate-lyase (R00472), glycine hydroxymethyltransferase (R00945), phosphoenolpyruvate carboxykinase (R00341) and malate dehydrogenase (R00342) in *M. extorquens* AM1 matches best with the submap containing the enzymes aconitase (R01324), isocitrate dehydrogenase (R00267), citrate synthase (R00351) and fumarate hydratase (R01082) in *E. coli*. Another part of the *M. extorquens* AM1 corresponding to enolase (R00658) and transaminase (R00585) matches well with that of pyruvate oxidation (R00209) and glyceraldehyde 3-phosphate dehydrogenase (R01061) in *E. coli*. The other elements of *M. extorquens* possess little consensus with *E. coli*.

data, which should be broken down into their original forms as in the KEGG to be consistent with other data.

COMMON USE CASES

The combination of data available from the CeCaFDB and the comparison functions via the web interface offer a powerful tool for structural comparisons and functional discovery of metabolic flux distribution. A few examples are provided below.

Flux Map Comparison

Clicking the alignment tab and selecting the flux distribution of *Methylobacterium extorquens* AM1 under mineral salts medium with methanol as the carbon source (45) and *Escherichia coli* under M9 minimal medium with glucose as the carbon source (46), and then performing an alignment based upon the 'Topology-based similarity' algorithm, produces a comparison result between the two flux distributions. The result includes a similarity score, a *P*-value and a URL linking to the graphical representation of the alignment solution. The alignment graph is directly displayed with a Cytoscape Web API (Figure 3). The olive line connecting the enzymes represents the matching relationship between the two flux maps. From Figure 3, it can be seen that the submap containing the enzymes L-Malate glyoxylate-lyase (R00472), glycine hydroxymethyltransferase (R00945), phosphoenolpyruvate carboxykinase (R00341) and malate dehydrogenase (R00342) in *M. extorquens* AM1 is the best match with the submap containing the enzymes aconitase (R01324), isocitrate dehydrogenase (R00267), citrate synthase (R00351) and fumarate hydratase (R01082) in *E. coli*. Another part of the *M. extorquens* AM1 corresponding to enolase (R00658) and transaminase (R00585) matches well with that of pyruvate oxidation (R00209) and glyceraldehyde 3-phosphate dehydrogenase (R01061) in *E. coli*. The other elements of *M. extorquens* possess little consensus with *E. coli*.

Interrelation between Flux Distributions

A key feature of the CeCaFDB is its extensive support for flux distribution queries against the selected part of the database, providing valuable hints regarding the interrelationships between them. Once the submitted flux distributions have been approved by administrators, users can perform similarity calculations on the submitted data using other data and can draw interrelations between all of the selected data. As an example, we conducted a correlation analysis on the flux distributions of *E. coli* under genetic and environmental manipulation. The correlation was analyzed for a total of 18 flux distributions from three references, with 10 of them being knocked out in different transcriptional factors or enzymes and 10 of them being cultivated under cultivated mediums and carbon sources (47–49). The detailed information for the flux distribution is shown in Supplementary Table S1. Supplementary Figure S2 indicates the hierarchy cluster tree of the ultimate results. Within the cluster tree, the FhlA and FadR mutants were located in the center and those under acetate or high glucose circumstances were isolated from the other cases. In the clustering process, the genetically and environmentally modified cases were intermixed. According to the cluster tree, we discovered that the distribution of the genetically modified flux is sparser than that from the changed environments, which points to a more stringent environmental effect on the flux distribution compared with the genetic manipulation on this occasion.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Dr Zhi Liang at the University of Science and Technology of China for his advice about the database.

FUNDING

Chinese National Natural Science Foundations [31200626, 31460233]; Programme for Changjiang Scholars and the Innovative Research Teams at the University [PCSIRT-1227]; Initial Fund for the Key Laboratory of Guizhou Province [2011-4005]; Guizhou Lianhe Foundation [LKS(2012)22]. Funding for open access charge: Chinese National Natural Science Foundations [31200626, 31460233]; Programme for Changjiang Scholars and the Innovative Research Teams at the University [PCSIRT-1227]; Initial Fund for the Key Laboratory of Guizhou Province [2011-4005]; Guizhou Lianhe Foundation [LKS(2012)22].
Conflict of interest statement. None declared.

REFERENCES

1. Sauer,U., Heinemann,M. and Zamboni,N. (2007) Genetics. Getting closer to the whole picture. *Science*, **316**, 550–551.
2. Sanford,K., Soucaille,P., Whited,G. and Chotani,G. (2002) Genomics to fluxomics and physiomics—pathway engineering. *Curr. Opin. Microbiol.*, **5**, 318–322.
3. Patti,G.J., Yanes,O. and Siuzdak,G. (2012) Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.*, **13**, 263–269.
4. Zhang,W., Li,F. and Nie,L. (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, **156**, 287–301.
5. Hovik,H., Yu,W.H., Olsen,I. and Chen,T. (2012) Comprehensive transcriptome analysis of the periodontopathogenic bacterium Porphyromonas gingivalis W83. *J. Bacteriol.*, **194**, 100–114.
6. Kang,H.J., Kawasawa,Y.I., Cheng,F., Zhu,Y., Xu,X., Li,M., Sousa,A.M., Pletikos,M., Meyer,K.A., Sedmak,G. *et al.* (2011) Spatio-temporal transcriptome of the human brain. *Nature*, **478**, 483–489.
7. Takeda,J., Suzuki,Y., Sakate,R., Sato,Y., Gojobori,T., Imanishi,T. and Sugano,S. (2010) H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Res.*, **38**, D86–D90.
8. Haug,K., Salek,R.M., Conesa,P., Hastings,J., de Matos,P., Rijnbeek,M., Mahendraker,T., Williams,M., Neumann,S., Rocca-Serra,P. *et al.* (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, **41**, D781–D786.
9. Wishart,D.S., Jewison,T., Guo,A.C., Wilson,M., Knox,C., Liu,Y., Djoumbou,Y., Mandal,R., Aziat,F., Dong,E. *et al.* (2013) HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–D807.
10. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C.A., Holland,T.A., Keseler,I.M., Kothari,A., Kubo,A. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.
11. Schellenberger,J., Park,J.O., Conrad,T.M. and Palsson,B.O. (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, **11**, 213.
12. Wiechert,W. (2001) 13C metabolic flux analysis. *Metab. Eng.*, **3**, 195–206.
13. Antoniewicz,M.R., Kelleher,J.K. and Stephanopoulos,G. (2007) Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab. Eng.*, **9**, 68–86.
14. Sauer,U. (2006) Metabolic networks in motion: 13C-based flux analysis. *Mol. Syst. Biol.*, **2**, 62.
15. Schmidt,K., Carlsen,M., Nielsen,J. and Villadsen,J. (1997) Modeling isotopomer distributions in biochemical networks using isotopomer mapping matrices. *Biotechnol. Bioeng.*, **55**, 831–840.
16. Wiechert,W., Schweissgut,O., Takanaga,H. and Frommer,W.B. (2007) Fluxomics: mass spectrometry versus quantitative imaging. *Curr. Opin. Plant Biol.*, **10**, 323–330.
17. Winter,G. and Kromer,J.O. (2013) Fluxomics—connecting 'omics analysis and phenotypes. *Environ. Microbiol.*, **15**, 1901–1916.
18. Covert,M.W., Xiao,N., Chen,T.J. and Karr,J.R. (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics*, **24**, 2044–2050.
19. Karr,J.R., Sanghvi,J.C., Macklin,D.N., Gutschow,M.V., Jacobs,J.M., Bolival,B. Jr, Assad-Garcia,N., Glass,J.I. and Covert,M.W. (2012) A whole-cell computational model predicts phenotype from genotype. *Cell*, **150**, 389–401.
20. Schatschneider,S., Huber,C., Neuweger,H., Watt,T.F., Puhler,A., Eisenreich,W., Wittmann,C., Niehaus,K. and Vorholter,F.J. (2014) Metabolic flux pattern of glucose utilization by Xanthomonas campestris pv. campestris: prevalent role of the Entner-Doudoroff pathway and minor fluxes through the pentose phosphate pathway and glycolysis. *Mol. Biosyst.*, **10**, 2663–2676.
21. Toya,Y., Nakahigashi,K., Tomita,M. and Shimizu,K. (2012) Metabolic regulation analysis of wild-type and arcA mutant Escherichia coli under nitrate conditions using different levels of omics data. *Mol. Biosyst.*, **8**, 2593–2604.
22. Niklas,J., Schrader,E., Sandig,V., Noll,T. and Heinzle,E. (2011) Quantitative characterization of metabolism and metabolic shifts during growth of the new human cell line AGE1.HN using time resolved metabolic flux analysis. *Bioprocess Biosyst. Eng.*, **34**, 533–545.
23. Crown,S.B., Indurthi,D.C., Ahn,W.S., Choi,J., Papoutsakis,E.T. and Antoniewicz,M.R. (2011) Resolving the TCA cycle and pentose-phosphate pathway of Clostridium acetobutylicum ATCC 824: Isotopomer analysis, in vitro activities and expression analysis. *Biotechnol. J.*, **6**, 300–305.
24. Kohlstedt,M., Becker,J. and Wittmann,C. (2010) Metabolic fluxes and beyond-systems biology understanding and engineering of microbial metabolism. *Appl. Microbiol. Biotechnol.*, **88**, 1065–1075.
25. Segre,D., Vitkup,D. and Church,G.M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 15112–15117.
26. Scott,D.A., Richardson,A.D., Filipp,F.V., Knutzen,C.A., Chiang,G.G., Ronai,Z.A., Osterman,A.L. and Smith,J.W. (2011) Comparative metabolic flux profiling of melanoma cell lines: beyond the Warburg effect. *J. Biol. Chem.*, **286**, 42626–42634.
27. Edwards,J.S., Ibarra,R.U. and Palsson,B.O. (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, **19**, 125–130.
28. Wendisch,V.F., de Graaf,A.A., Sahm,H. and Eikmanns,B.J. (2000) Quantitative determination of metabolic fluxes during coutilization of two carbon sources: comparative analyses with Corynebacterium glutamicum during growth on acetate and/or glucose. *J. Bacteriol.*, **182**, 3088–3096.
29. Blank,L.M., Kuepfer,L. and Sauer,U. (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.*, **6**, R49.
30. Schuetz,R., Zamboni,N., Zampieri,M., Heinemann,M. and Sauer,U. (2012) Multidimensional optimality of microbial metabolism. *Science*, **336**, 601–604.
31. Nogales,J., Gudmundsson,S., Knight,E.M., Palsson,B.O. and Thiele,I. (2012) Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 2678–2683.
32. Shlomi,T., Benyamini,T., Gottlieb,E., Sharan,R. and Ruppin,E. (2011) Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS Comput. Biol.*, **7**, e1002018.
33. Crown,S.B. and Antoniewicz,M.R. (2013) Publishing 13C metabolic flux analysis studies: a review and future perspectives. *Metab. Eng.*, **20**, 42–48.
34. Suthers,P.F., Burgard,A.P., Dasika,M.S., Nowroozi,F., Van Dien,S., Keasling,J.D. and Maranas,C.D. (2007) Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metab. Eng.*, **9**, 387–405.
35. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
36. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
37. Cui,Q., Lewis,I.A., Hegeman,A.D., Anderson,M.E., Li,J., Schulte,C.F., Westler,W.M., Eghbalnia,H.R., Sussman,M.R. and Markley,J.L. (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.*, **26**, 162–164.
38. Stelling,J., Klamt,S., Bettenbrock,K., Schuster,S. and Gilles,E.D. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190–193.
39. Pinter,R.Y., Rokhlenko,O., Yeger-Lotem,E. and Ziv-Ukelson,M. (2005) Alignment of metabolic pathways. *Bioinformatics*, **21**, 3401–3408.
40. Zhenping,L., Zhang,S., Wang,Y., Zhang,X.S. and Chen,L. (2007) Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, **23**, 1631–1639.
41. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
42. Ogata,H., Fujibuchi,W., Goto,S. and Kanehisa,M. (2000) A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, **28**, 4021–4028.

43. Kelley,B.P., Yuan,B., Lewitter,F., Sharan,R., Stockwell,B.R. and Ideker,T. (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.*, **32**, W83–W88.
44. Rohn,H., Hartmann,A., Junker,A., Junker,B.H. and Schreiber,F. (2012) FluxMap: a VANTED add-on for the visual exploration of flux distributions in biological networks. *BMC Syst. Biol.*, **6**, 33.
45. Van Dien,S.J., Strovas,T. and Lidstrom,M.E. (2003) Quantification of central metabolic fluxes in the facultative methylotroph methylobacterium extorquens AM1 using 13C-label tracing and mass spectrometry. *Biotechnol. Bioeng.*, **84**, 45–55.
46. Bianco,C., Imperlini,E., Calogero,R., Senatore,B., Pucci,P. and Defez,R. (2006) Indole-3-acetic acid regulates the central metabolic pathways in Escherichia coli. *Microbiology*, **152**, 2421–2431.
47. Blank,L.M., Kuepfer,L. and Sauer,U. (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.*, **6**, R49.
48. Lemuth,K., Hardiman,T., Winter,S., Pfeiffer,D., Keller,M.A., Lange,S., Reuss,M., Schmid,R.D. and Siemann-Herzberg,M. (2008) Global transcription and metabolic flux analysis of Escherichia coli in glucose-limited fed-batch cultivations. *Appl. Environ. Microbiol.*, **74**, 7002–7015.
49. Zhao,J. and Shimizu,K. (2003) Metabolic flux analysis of Escherichia coli K12 grown on 13C-labeled acetate and glucose using GC-MS and powerful flux calculation method. *J. Biotechnol.*, **101**, 101–117.